Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces

Patrick Ehlen, Michael F. Schober

Department of Psychology
The New School for Social Research
{ehlenp, schober}@newschool.edu

Frederick G. Conrad

Survey Research Center
Institute for Social Research, University of Michigan
FConrad@isr.umich.edu

Abstract

Computer-based interviewing systems could use models of respondent disfluency behaviors to predict a need for clarification of terms in survey questions. We compare simulated speech interfaces that use two such models—a generic model and a stereotyped model that distinguishes between the speech of younger and older speakers—to several non-modeling speech interfaces in a task where respondents provided answers to survey questions from fictional scenarios. Our modeling procedure found that the best predictor of conceptual misalignment was a critical *Goldilocks range* for response latency, outside of which responses are more likely to be conceptually misaligned. Different Goldilocks ranges are effective for younger and older speakers.

Keywords: conceptual alignment, misalignment cues, Goldilocks range, stereotyped modeling, speech survey interfaces

1 Introduction

Let's say you're cooking spaghetti in the kitchen and the phone rings and before you know it you've agreed to answer some questions in a survey interview. The interviewer asks you, "How many hours per week do you usually work at your job?" As a corporate lawyer who aspires to be a professional chef, your answer is potentially elaborate. You get paid for making calls and eating lunch and playing golf, and you might say you're working even now, practicing for your next career. But instead of going into all that, you find yourself saying, "Well… uh… usually seventy."

What are you up to? You've made a guess about what you think the interviewer is getting at, while recognizing that your own concept of "work" or "job" or "usually" might be different. That is, you've recognized the possibility of *conceptual misalignment* (Schober, 1998). And instead of giving an elaborate answer that would defy Grice's (1975) cooperative principle—which sees to it that we don't say much more about things than we think we need to for the task at hand—you answer based on that guess. But at the same time, you've laid your own awareness of possible conceptual misalignment out on the table for the interviewer to note, using subtle signals that indicate things may not be so straightforward: you hedge ("well"), you pause ("…"), and you insert a filler ("uh") before you answer, all of which are disfluent behaviors that can signal when a speaker has a problem (Brennan & Williams, 1995; Schober & Bloom, 2004; Smith & Clark, 1993). We'll call these signals *misalignment cues*.

So communication happens on two planes: on one plane you answer the question you think is being asked, providing information that attends to the task at hand, while on another plane you send out some misalignment cues that manage possible subtasks of the dialogue, signaling misalignment information that may or may not be so important but is subtly offered in case it proves relevant (Bangerter & Clark, 2003). Now it's up to the interviewer to catch on to those signals and decide if she wants to pursue that misalignment problem further.

But the interviewer's recognition of your misalignment problem is by no means assured, partly because people speak disfluently for all kinds of reasons, and partly because she brings her own expectations to the dialogue about what that disfluency might mean. If you're very far into the interview and you've said "well" and "uh" in every answer, by now the interviewer may have decided you're just a person who says "well" and "uh" a lot, and may not infer any misalignment at all. That is, she creates an

individualized model of your disfluency behaviors, and may only infer misalignment when those behaviors deviate from her expectations according to that model. Or, even if this is the first question in the interview, she might bring some established assumptions—a *stereotyped model*—that shape her expectations about your speech. Did we mention yet that you're 76 years old? And since older speakers are known to be more disfluent (Schow et al., 1978; Shrivastav et al., 2003; Yairi & Clifton, 1972), and interviewers act differently with older respondents compared to younger ones (Bradburn et al., 1979), she may be working from a model that tells her that your hedging and pausing and fillering do not indicate any misalignment problem at all, unless your use of those cues deviates from the stereotyped model that shapes her idea of how seventy-something-year-old people should speak.

In short, we can speculate that misalignment cues are overdetermined and that, like words in general, their interpretation requires a dynamic and interactive process, inasmuch as that interpretation depends on the expectations—or speaker models (Schober & Brennan, 2003)—that people bring to a dialogue.

This overdetermination poses a problem for anyone who might have the idea of using these misalignment cues to try to automatically diagnose conceptual misalignment problems in survey interviews. Could a machine algorithm listen for certain misalignment cues and then offer clarification of key words like "work" or "job" when it hears those cues? Or would it be better if that machine were also tuned in to other aspects of the interaction that people often pay attention to, such as the age of the person it's talking with (using a stereotyped model of misalignment cues), and changed its expectations accordingly?

After all, telephone interviewers of the future may well consist of machine speech interfaces that recognize respondents' unconstrained speech, and working dialogue systems that conduct interviews are already in use (Blyth, 1997; Cole et al., 1994). But if they are to be as effective as human interviewers, these machines would need to detect and predict cases of possible conceptual misalignment, where respondents need clarification of concepts in a question. Clarification has been shown to improve comprehension—and thus data quality—in survey interviews (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad & Fricker, 2004). Ideally, that clarification would be targeted and offered only when necessary, without being an intrusive burden, perhaps by modeling expectations about the speaker's needs. Our purpose here is to explore how that process of modeling and targeting might work in automated speech survey interfaces, comparing the effectiveness of a generic model of misalignment cues (that treats all respondents the same) to a stereotyped model (that distinguishes respondents by age group), and also comparing these modeling procedures to interfaces that rely on other criteria to address problems of conceptual misalignment.

2 Experimental Method and Results

To identify possible misalignment cues in answers to survey questions and derive models that could then be tested for their effectiveness at identifying cases of conceptual misalignment, this study was divided into two phases: (1) a *model-derivation phase* that collected data that were used to create and validate generic and stereotyped models, and (2) a *model-implementation phase* that tested implementations of the models derived from the first phase on a new set of respondents. These phases consisted of five conditions: three in the model-derivation phase, and two in the model-implementation phase. Since all five conditions were similar, we'll first look at the overall design, and then at the methods and results of each of the two phases.

2.1 Overall Design

Participants completed a referential communication survey task similar to the design used by Schober and Conrad (1997), answering a selected set of twelve survey questions from existing, ongoing U.S. government surveys. Rather than responding about their own lives, they responded about fictional scenarios so the accuracy of their answers could be determined. And rather than being interviewed by a person, participants answered questions presented by a computer over a standard telephone. All participants were asked the same questions, though scenarios varied in terms of how well potentially-ambiguous concepts in the scenarios mapped onto the questions asked. The method by which participants received clarification about those potentially-ambiguous concepts also varied, constituting five experimental conditions: (1) a baseline, *no clarification* condition where no clarification was provided, (2) a *respondent-initiated clarification* condition where clarification was provided only when explicitly requested, (3) a *required clarification* control and validation condition, where clarification was provided

for every question, (4) a *generic respondent model* condition where clarification was provided on request and also provided automatically from a "generic respondent misalignment" model, and (5) a *stereotyped respondent model* condition where clarification was provided on request and also from a "stereotyped respondent misalignment" model that distinguished between older and younger respondents. In the model-derivation phase, measures of potential misalignment cues taken from responses in the *no clarification* condition and the *respondent-initiated clarification* condition were used to derive models developed for the two modeling conditions, and measures from responses collected in the *required clarification* condition were used to validate those models. In the model-implementation phase, these two models were then tested using the *generic respondent model* and *stereotyped respondent model* conditions.

Participants. 100 participants were divided into two age groups and randomly assigned across the five conditions, with 10 participants per group per condition. Participants were assigned to groups by age: one group of 50 participants over the age of 65 constituted an "older" group, and the other group of 50, younger than age 40, constituted a "younger" group. Genders were balanced across conditions, with 5 males and 5 females from each age group participating in each condition, and all were native English speakers.

Questions and Scenarios. Each participant was asked to answer twelve survey questions taken from ongoing survey interviews conducted by the US Bureau of Labor Statistics (BLS) in three different domains: Four questions about housing came from the Consumer Price Housing Index survey (e.g., "How many people live in this house?"), four questions about work situations came from the Current Population Survey (e.g., "How many hours per week do you usually work at your job?"), and four questions about purchases came from the Current Point of Purchase Survey (e.g., "Have you purchased or had expenses for household furniture?"), all of which have been used in earlier studies (e.g., Schober & Conrad, 1997). For each question, survey designers at BLS had already developed official definitions for key concepts designed to clarify whether, for example, a "floor lamp" should be counted as a piece of "household furniture," which helped us to infer whether a participant's understanding of the key concept in a question matched the intended definition—that is, it allowed us to measure conceptual alignment between the participant and the question designer.

The official definitions sounded something like this: *"Let me give you our definition of household furniture. Include tables, chairs, footstools, sofas, china cabinets, utility carts, bars, room dividers, bookcases, desks, beds, mattresses, box springs, chests of drawers, night tables, wardrobes, and unfinished furniture. Do not include TV, radio, and other sound equipment, lamps and lighting fixtures, outdoor furniture, infants' furniture, or appliances."*

Participants answered these questions while looking at fictional scenarios in which the key concepts were manipulated so the response accuracy of participants' answers could be used to infer conceptual alignment. Each question had two alternate scenarios: either a *straightforward mapping* where the description in the scenario mapped onto the key concept of the question in a clear and simple way, or a *complicated mapping* where the description in the scenario mapped onto the question in a way that was more open to interpretation, making it hard to answer correctly without information about the official definition of the key concept. For example, a participant who is asked "Has Kelly purchased or had expenses for household furniture?" might see a straightforward scenario that showed Kelly's receipt for the purchase of an end table—which most people would interpret as household furniture—or instead might see a complicated scenario with a receipt for a floor lamp, which some may interpret as household furniture, even though the official definition states that floor lamps should not be counted. Half of the scenarios seen by each participant were straightforward mappings, and half were complicated mappings.

Procedure. Participants sat at a desk in a laboratory with a packet of scenarios and a regular Bell-style telephone and were asked to dial a number when they felt familiar enough with the scenarios to answer questions about them. Though participants were told they would be interviewed by a computer, the number they dialed was actually answered by a computer in the next room that was controlled by an experimenter, who used a Wizard-of-Oz telephony interface to present the questions and reply to the participants' answers using a synthesized text-to-speech voice.

In the *no clarification* condition, participants heard a question, provided an answer, and immediately moved on to the next question. In the *respondent-initiated clarification* condition, participants were told

they could ask for clarification of terms if they felt like they needed it. If they asked for clarification, the synthesized voice read them the official definition of the key concept, and then asked the question again. In the *required clarification* condition, clarification was provided for every question. After participants answered a question they automatically heard the scripted definition and were presented with the question a second time, allowing them to change their answers if they wished. Finally, in the *generic respondent model* and the *stereotyped respondent model* conditions, respondents answered the question and were then given clarification if the misalignment cues in their initial answers met the criteria of the respondent models for that condition, followed by a chance to answer the question a second time.

2.2 Model-Derivation Phase

The model-derivation phase collected the data needed to derive and validate the models used in the modeling conditions of the second phase. It consisted of the *no clarification,* the *respondent-initiated clarification*, and the *required clarification* conditions. Data from the *no clarification* and *respondent-initiated clarification* conditions were used to create models of the behaviors that best predicted conceptual misalignment, and the *required clarification* condition was used to validate those models.

Predictors Used. To derive the models, the following misalignment cues immediately following each question were pooled from responses in the *no clarification* and the *respondent-initiated clarification* conditions and tallied for use as potential predictors: length of *response latency* in milliseconds from the end of the question to the beginning of an answer, *fillers*, *hedges*, *restarts*, *repeats*, *repairs*, *reports*, and *mutters*. The criteria for coding these behaviors were adapted from Bortfeld et al. (2001).

One other potential cue was added early in the experiment after participants were seen performing a consistent behavior we had not considered: While most answers were straightforward and unelaborated (such as "yes" or "fifty"), some responses also repeated a word or words from the question that had just been asked that often included the key concept word(s). This behavior can be seen as a joint action (à la Clark, 1994) that "picks up" any potentially-ambiguous words from a question and "keeps them in play" in the dialogue, allowing them to be confirmed and negotiated further by both parties.

Consider the question, "How many hours per week does Mindy usually work at her job?" A response like "fifty" does not invite any further negotiation of the terms of the question, and does not offer any recognition that the word "usually" is open to interpretation. However, a response like, "Usually, fifty" picks up the term "usually" as a way of keeping it in play so it can be confirmed or negotiated. Such behavior may show some awareness for the respondent that a concept is open to interpretation. We could call this behavior a *referential confirmation pick-up*; or more simply, a *confirmation pick-up*.

The cues mentioned above (response latency, fillers, hedges, restarts, repeats, repairs, reports, mutters, and confirmation pick-ups) were all used as predictors to derive models of misalignment profiles for both the generic and the stereotyped modeling conditions, using an ordinary least squares regression to determine the best predictors of conceptual misalignment, inferred through response accuracy *(A)* as the criterion variable. The regression equation for each model began with factors for *response latency (L), fillers (F), hedges (H), restarts (RS), repeats (RP), repairs (RA), reports (RO), clarification requests (CR), repeat requests (RR), confirmation pick-ups (CP),* and *mutters (M)*, creating the following least-squares equation:

$$A = \beta_1 L + \beta_2 F + \beta_3 H + \beta_4 RS + \beta_5 RP + \beta_6 RA + \beta_7 RO + \beta_8 CR + \beta_9 RR + \beta_{10} CP + \beta_{11} M + e$$

Coefficients (β) for each of these potential cues were determined that yielded the smallest residual constant (e) and factors were eliminated using backward variable elimination.

Age Group Differences in Disfluency Rates. The first question we asked was whether there were indeed significant differences in disfluency rates between the two age groups. Findings from the first two conditions reveal that older participants were in fact more disfluent than younger ones, in keeping with previous findings (Bortfeld et al., 2001; Schow et al., 1978; Shrivastav et al., 2003; Yairi & Clifton, 1972). When compared to the younger group, the older group provided significantly more fillers ($F(1,478) = 6.47$, $p = .011$), restarts ($F(1,478) = 6.13$, $p = .014$), repeats ($F(1,478) = 20.55$, $p < .001$), repairs ($F(1,478) = 5.47$, $p = .02$), reports ($F(1,478) = 15.50$, $p < .001$), repeat requests ($F(1,478) = 10.35$, $p = .001$), and mutters ($F(1,478) = 9.31$, $p = .001$).

Response Latency and the Goldilocks Ranges. Against expectation, the raw measure of response latency initially did not show any correlation with accuracy (e.g., longer latencies predicting greater inaccuracy), and it did not show significant differences between age groups. A closer look at the data suggested that responses to complicated mappings that were too fast were often as likely to lead to inaccurate responses as responses that were too slow, leading to a critical range of latency that is "just right" for predicting when people may provide an accurate response, but outside of which they are more likely to provide an inaccurate one. That range can be thought of as a *Goldilocks range* for response latency as a predictor of accuracy.

Goldilocks ranges for the models were derived and tested by finding the latency ranges that yielded a maximal adjusted R^2 for each first-pass regression model using all predictors. When our initial response latency predictors were replaced with variables that specified whether the latency of a response fell outside the Goldilocks ranges, these models yielded better predictions of accuracy than a single-threshold generic model (adjusted R^2 of -.041, p = .610) in both the generic (adjusted R^2 of .003, p = .454) and stereotyped (adjusted R^2 of .212, p = .048) cases.

The Goldilocks range determined from responses from all participants without regard to age group yielded a range between 2 and 7.2 seconds, which is a range that can be used to help derive a generic model. Analysis of the two age groups using two independent Goldilocks factors (one for younger respondents and another for older) revealed a Goldilocks range of 4.6 to 10.2 seconds for the younger group, and 2.6 to 4.35 seconds for the older group. While these two stereotyped ranges bring to light a clear difference between how the two groups answer questions, they do not support the prediction that "older people take longer to answer" in general. Rather, the response latency range in which older people can be expected to provide a conceptually aligned answer is attenuated, and also shifted to a much *faster* response time than we see for the younger group (perhaps indicating that older respondents apply a different kind of knowledge or answering strategy).

Generic and Stereotyped Models. Multiple-pass regression analyses of all potential predictors showed that in fact a participant's failure to respond within the response latency Goldilocks range proved the single significant predictor of inaccurate responses for both the generic model and the stereotyped model. Confirmation pick-ups surfaced as the second-most enduring predictor in both regression models, though they predicted *accurate* responses rather than *inaccurate* ones. At p = .15 in the generic and p = .23 in the stereotyped model, confirmation pick-ups did not reach the criterion of significance to be included in the final regression, though a repeated measures analysis showed confirmation pick-ups as predictive of *accurate* answers for older respondents, $F(1,38)=5.28$, p = .027, and also predictive of complicated mappings for both age groups, $F(1,38)=4.90$, p = .033.

So our models of both the generic and the stereotyped cases used only the Goldilocks range factors to predict conceptual misalignment. Responses that fell within that range were more likely to be aligned and did not warrant offers of clarification, while responses falling outside the range were more likely to be misaligned, and therefore warranted clarification.

Validation of Models. Because the *required clarification* condition solicited two responses from participants—one that came before the participant heard clarification and one that came after—this condition provided an ideal platform to test our models on an independent data set, since we knew how respondents would have answered both with and without clarification. For the generic model, when participants made an inaccurate response, the model predicted 53.3% of these as responses that came either too slow or too fast. The stereotyped model predicted 83.3% of participants' inaccurate responses. But would these validation results extend to actual implementation of the models? That question could only be answered experimentally.

2.3 Model-Implementation Phase

The model-implementation phase implemented and tested the derived generic and stereotyped models. It consisted of the *generic respondent model* and the *stereotyped respondent model* conditions.

Response Accuracy. Do models that use generic or stereotyped Goldilocks ranges actually help to reduce conceptual misalignment and improve response accuracy when compared to the other non-modeled conditions? The short answer is that they do: both modeling conditions result in significantly higher

accuracy on complicated mappings than not modeling at all, with modeling conditions showing accuracy ratings that are reliably on-par with providing clarification for every question.

While overall mean accuracy for complicated mappings was only 52%, the differences in response accuracy for complicated mappings varied greatly by condition, as shown in Figure 1. Though respondents reliably differed in response accuracy by condition [$F1(4,95) = 35.87$, $p < .001$; $F2(4,55) = 19.40$, $p < .001$], this effect shows a significant difference only between the *respondent-initiated clarification* condition and the *generic respondent model* condition. With no clarification at all, accuracy on complicated mappings reached only 20%. When participants were allowed to ask for clarification, accuracy rose to a somewhat higher 28%, $F1(1,95) = 1.46$, $p = n.s.$; $F2(1,55) = .80$, $p = n.s.$

But for the generic modeling condition where all participants were offered clarification according to the same Goldilocks range, accuracy on complicated mappings reached a reliably higher 64%, $F1(1,90) = 35.03$, $p < .001$; $F2(1,55) = 19.09$, $p < .001$. And when different Goldilocks ranges were tailored for the respective younger and older groups, accuracy in the stereotyped condition reached 72%, $F1(1,90) = 1.46$, $p = n.s.$; $F2(1,55) = .80$, $p = n.s.$ The highest accuracy on complicated mappings came when all participants heard clarification every time, at 77%, which was not reliably different from accuracy in the stereotyped condition, $F1(1,90) = .65$, $p = n.s.$; $F2(1,55) = .35$, $p = n.s.$ Although these accuracy rates for the *generic respondent model*, the *stereotyped respondent model*, and the *required clarification* conditions do not differ significantly from each other, they do reflect a significant linear trend, $F1(1,90) = 128.10$, $p < .001$; $F2(1,55) = 68.89$, $p < .001$, in which the higher rates of clarification afforded by increasingly fine-tuned modeling (or from providing clarification for every question) lead to higher accuracy.

Not surprisingly, and consistent with prior findings (Schober & Conrad, 1997; Schober, Conrad & Fricker, 2004), participants fared very well overall at answering questions to straightforward mapping scenarios (mean accuracy of 94%), which varied little across conditions and age groups.

While stereotyped modeling did not lead to reliably better overall response accuracy on complicated mappings than generic modeling, there was a marginal difference by age group by participant, $F1(1,90) = 3.49$; $F2(1,110) = 2.58$, $p = n.s.$, where older respondents fared marginally better from stereotyped modeling (see Figure **2**), reaching an accuracy of 75% as compared to only 57% in the generic modeling condition, $F1(1,45) = 3.79$, $p = .058$; $F2(1,55) = 3.79$, $p = .057$. Younger participants, however, fared about the same with either type of modeling, reaching 68% accuracy with stereotyped modeling compared to 72% with generic modeling, $F1(1,45) = .17$, $p = n.s.$; $F2(1,45) = .09$; $p = n.s.$ There was no reliable interaction between age group and condition for all conditions $F1(4,90) = .98$, $p = n.s.$; $F2(4,110) = .73$, $p = n.s.$ Older participants did not fare reliably better from receiving clarification every time (in fact they appear slightly worse) than they did from stereotyped modeling, with an accuracy of 70%, $F1(1,45) = .28$, $p = n.s.$; $F2(1,55) = .28$, $p = n.s.$ And although younger participants fared best in the required clarification condition, with an accuracy of 83%, this increase in accuracy was not reliably better than accuracy in the *generic respondent model*, $F1(1,45) = 2.08$, $p = n.s.$; $F2(1,55) = 1.14$, $p = n.s.$, or the *stereotyped respondent model*, $F1(1,45) = 3.43$, $p = n.s.$; $F2(1,55) = 1.89$, $p = n.s.$

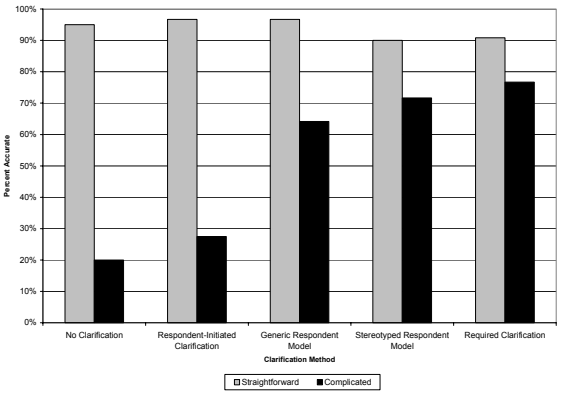

Figure 1: Response accuracy by condition for all ages

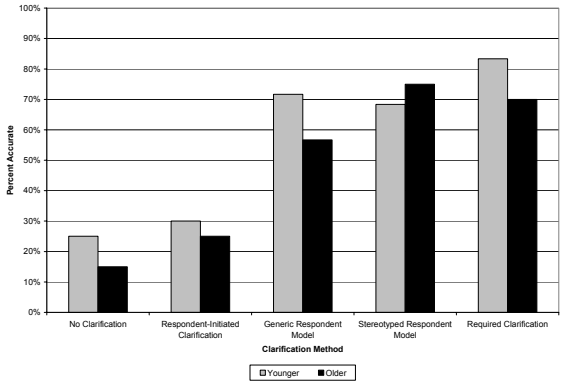

Figure 2: Response accuracy by age on complicated mappings

Time to Answer Each Question. How much time will be added to an interview if these various methods of providing clarification are implemented? The general trend was that the more clarification was given, the

more time it took each participant to get through each question. Question-response sequences went fastest when little clarification was given, in the *no clarification* condition (16.1 seconds for the younger group and 19.2 seconds for the older group) and the *respondent-initiated clarification* condition (younger = 15.0 secs, older = 20.2 secs). More time was needed per question in the *generic respondent model* condition (younger = 39.4 secs, older = 42.5 secs) and slightly more in the *stereotyped respondent model* condition (younger = 50.0 secs, older = 51.2 secs). The most time was needed in the *required clarification* condition (younger = 60.9 secs, older = 66.1 secs). These differences were reliable by condition, $F(4, 84) = 173.76$, $p < .001$, and by age group, $F(1,84) = 6.95$, $p = .01$, with no significant interaction between the two, $F(4,84) = .31$; $p = n.s.$ As seen in Figure 3, the younger group is consistently faster. Respondents took reliably less time to answer questions under the *stereotyped respondent model* method than they did under the *required clarification* method, $F(1,35) = 36.84$, $p < .001$. This difference is worth noting, since the earlier comparison of these two conditions showed no reliable difference in *accuracy*. Respondents who received stereotyped modeling were just as accurate as respondents who received clarification for every question, but spent significantly less time on each question.

Respondent Satisfaction Ratings. Regardless of the method used to provide clarification (whether because of modeling or simply provided for every question), respondents did not reliably prefer one method over another. Overall mean satisfaction was 5.41, with a marginally higher rating coming from older respondents (mean of 5.73) than younger ones (mean of 5.08), showing no reliable effect by age group ($F(1,54) = 2.11$, $p = n.s.$) or condition ($F(2,54) = .49$, $p = n.s.$). Satisfaction results are shown in Figure 4.

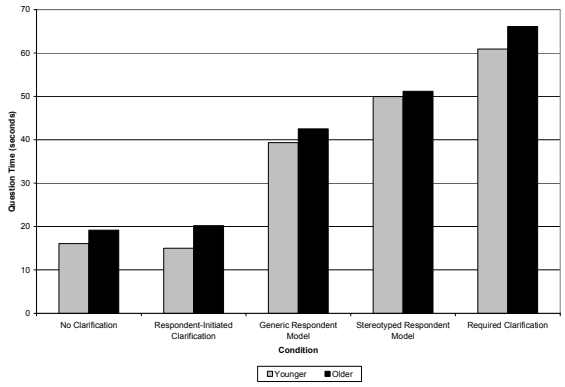

Figure 3: Average time to answer each question

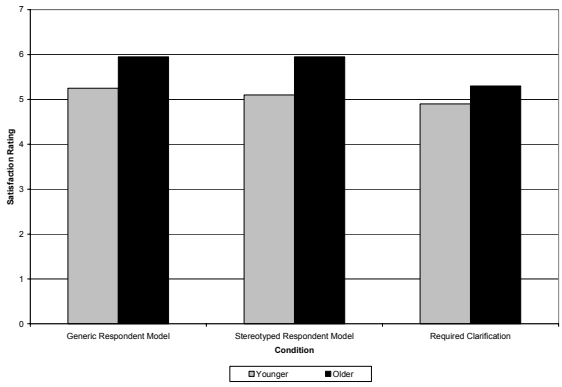

Figure 4: Respondent satisfaction with clarification

3 Conclusion

Our purpose here was to look into exploiting that "second plane" of communication in which signals like misalignment cues help to manage the dialogue process, and to see if generic or stereotyped models of those signals could help diagnose cases of conceptual misalignment in an automated survey system. After deriving both generic and stereotyped models of various misalignment cues, we found the best cue is conspicuous in its absence: a critical *Goldilocks range* in response latency, outside of which people are more likely to give conceptually misaligned responses. A generic Goldilocks range of 2 to 7.2 seconds is effective at predicting overall levels of conceptual alignment. But that range is more effective when tailored for older and younger speakers, where older speakers are less likely to provide a conceptually aligned response if their answers do not fall within a range that is *briefer* and slightly *sooner* (2.6 to 4.35 seconds) than the range used by younger speakers (4.6 to 10.2 seconds).

In a nutshell, "help helps—and tailored help helps better." Whatever the modeling method, both models led to more offers of clarification, and more clarification led to greater accuracy. Whether they received this clarification as a result of modeling through the Goldilocks ranges or simply received it for every question, respondents were reliably more accurate when given clarification automatically than when they were left to ask for it on their own or were not given clarification at all. So help helps, but tailored help helps better, since both older and younger age groups were just as accurate under stereotyped modeling as when they received clarification every time, yet took reliably *less time* to answer

questions in the stereotyped modeling condition. In addition, they were just as satisfied with the clarification they received under stereotyped modeling as they were with receiving it every time.

So it seems expectations about the speech of different types of respondents can be successfully modeled to help assess problematic answers to questions or conceptual misalignment in general. And computer interviewing systems could be designed to permit flexible interactions that help respondents provide data that are in line with the intentions of the question authors.

On a final note, while older speakers showed more disfluent behaviors than younger speakers when verbally answering questions asked by a computer interviewer—using reliably more fillers, restarts, repeats, repairs, reports, repeat requests, and mutters—most of these other disfluent behaviors were not highly significant predictors in regression models that sought to identify cases of conceptual misalignment. Does this mean that these other misalignment cues would not be effective in models that seek to predict conceptual misalignment? On the contrary, these other cues could prove much more effective if modeled on an individualized, case-by-case basis. Now that we've seen that some misalignment cues are in fact relative and overdetermined under stereotyped modeling, a real test of that idea calls for an interactive and highly-dynamic approach that creates and adjusts its model of a person's style of speech—that is, working from an *individualized model*—even while that person is still speaking.

4 References

Bangerter, A., & Clark, H.H. (2003). Navigating joint projects with dialogue. *Cognitive Science, 27*, 195-225.

Blyth, W.G. (1997). Developing a speech recognition application for survey research. In L.E. Lyberg, P.P. Biemer, M. Collins, E. de Leeuw, C.S. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 249-266). New York: Wiley.

Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F., & Brennan, S.E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*, 123-147.

Bradburn, N., Sudman, S., & Associates. (1979). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.

Brennan, S.E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34*, 383-398.

Clark, H.H. (1994). Managing problems in speaking. *Speech Communication, 15*, 243-250.

Cole, R.A., Novick, D.G., Fanty, M., Vermeulen, P.J.E., Sutton, S., & Burnett, D. (1994) A prototype voice-response questionnaire for the U.S. Census. In *Proceedings of the International Conference on Speech and Language Processing* (pp. 683-686). ICSLP.

Conrad, F.G., & Schober, M.F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*, 1-28.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 225-242). New York: Seminar Press.

Schober, M.F. (1998). Conversational evidence for rethinking meaning. *Social Research, 65*, 511-534.

Schober, M.F., & Brennan, S.E. (2003). Processes of interactive spoken discourse: The role of the partner. In A.C. Graesser, M.A. Gernsbacher, & S.R. Goldman (Eds.), *Handbook of Discourse Processes* (pp. 123-164). Mahwah, NJ: Lawrence Erlbaum Associates.

Schober, M.F., & Bloom, J.E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse Processes, 38*, 287-308.

Schober, M.F., & Conrad, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61,* 576-602.

Schober, M.F., Conrad, F.G., & Fricker, S.S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 17*.

Schow, R.L., Christensen, J.M., Hutchinson, J.M., & Nerbonne, M.A. (1978). *Communication disorders of the aged*. Baltimore: University Park Press.

Shrivastav, R., Hollien, H., Brown, Jr., W.S., Rothman, H.B., & Harnsberger, J.D. (2003). Shifting perceptions of age in voice. Poster presented at the *146th Meeting of the Acoustical Society of America, Nov. 10-14, 2003, Austin, TX*.

Smith, V.L., & Clark, H.H. (1993). On the course of answering questions. *Journal of Memory and Language, 32*, 25-38.

Yairi, E., & Clifton, Jr., N.F. (1972). Disfluent speech behavior of preschool children, high school seniors and geriatric persons. *Journal of Speech and Hearing Research, 15*, 714-719.